AD-A035 732 AIR FORCE HUMAN RESOURCES LAB BROOKS AFB TEX F/6 5/9
EFFECTS OF ITEM-OPTION WEIGHTING ON THE RELIABILITY AND VALIDIT--ETC(U)
DEC 76 M J REE

UNCLASSIFIED AFHRL-TR-76-76 NL

1 OF 1
AD-A
035 732

END
DATE
FILMED
3-25-77
NTIS

EFFECTS OF ITEM-OPTION WEIGHTING ON THE RELIABILITY AND VALIDITY OF THE AFOQT FOR PILOT SELECTION

AIR FORCE HUMAN RESOURCES LABORATORY
BROOKS AIR FORCE BASE, TEXAS

DECEMBER 1976

AFHRL-TR-76-76

AIR FORCE HUMAN RESOURCES LABORATORY

EFFECTS OF ITEM-OPTION WEIGHTING ON THE RELIABILITY AND VALIDITY OF THE AFOQT FOR PILOT SELECTION

By

Malcolm James Ree

PERSONNEL RESEARCH DIVISION
Lackland Air Force Base, Texas 78236

December 1976
Final Report for Period March 1975 – September 1976

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

NOTICE

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This final report was submitted by Personnel Research Division, Air Force Human Resources Laboratory, Lackland Air Force Base, Texas 78236, under project 7719, with HQ Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base, Texas 78235.

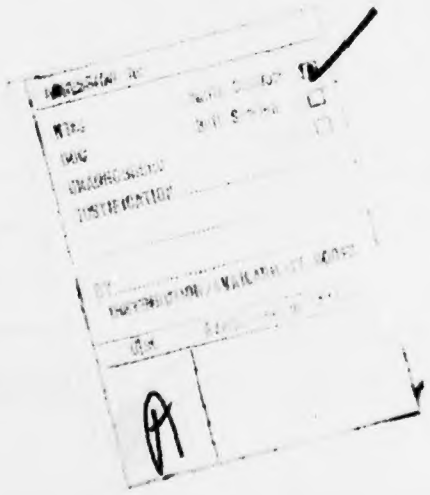
This report has been reviewed and cleared for open publication and/or public release by the appropriate Office of Information (OI) in accordance with AFR 190-17 and DoDD 5230.9. There is no objection to unlimited distribution of this report to the public at large, or by DDC to the National Technical Information Service (NTIS).

This technical report has been reviewed and is approved.

LELAND D. BROKAW, Technical Director
Personnel Research Division

Approved for publication.

DAN D. FULGHAM, Colonel, USAF
Commander



REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFHRL-TR-76-76	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): EFFECTS OF ITEM-OPTION WEIGHTING ON THE RELIABILITY AND VALIDITY OF THE AFOQT FOR PILOT SELECTION	5. TYPE OF REPORT & PERIOD COVERED: Final, March 1975 – September 1976
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Malcolm James Ree	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Personnel Research Division, Air Force Human Resources Laboratory, Lackland Air Force Base, Texas 78236	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 62703F, 77191221
11. CONTROLLING OFFICE NAME AND ADDRESS: HQ Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base, Texas 78235	12. REPORT DATE: December 1976
	13. NUMBER OF PAGES: 12
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

item analysis
least squares
pilot selection
reliability
test construction
validity

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This study was designed to investigate the effects of the use of differential item-option weights on the reliability and validity of the Air Force Officer Qualifying Test (AFOQT) as used for pilot selection. Two groups of subjects were selected from a pool of 3,400 students who had been admitted to undergraduate pilot training from 1969 to 1972. Using an extension of the method attributed to Guttman, item-option weights were generated and used to score the AFOQT on a cross-validation group. Corrected-for-guessing scores were computed for comparative purposes. The internal consistency reliability of the subtests and of the pilot composite increased with item-option weights were applied.

Item 20 (Continued)

The validity of most subtests and the Pilot Composite was higher for item-option weighted scores than for corrected-for-guessing scores.

The reliability of the item-option weights was moderate.

PREFACE

This research was conducted under project 7719; Air Force Development of Selection, Assignment, Performance Evaluation, Retention, and Utilization Devices; task 771912, Armed Forces Programs.

The professional and technical assistance provided during the course of this research by the Computational Sciences Division is greatly appreciated.

TABLE OF CONTENTS

LIST OF TABLES

EFFECTS OF ITEM-OPTION WEIGHTING ON THE RELIABILITY AND VALIDITY OF THE AFOQT FOR PILOT SELECTION

I. INTRODUCTION

The concept of differentially weighting test items or the answers within an item spans the history of psychometric research. It is logically appealing to assign a greater weight to the answers of more discriminating test questions.

Guttman (1941) proposed a method of scaling items to maximize the squared correlation ratio, Eta^2, between item-option categories and some criterion of interest. This method has been adapted to right-wrong scored items and has been investigated for some time. The weights assigned to item options by this method are the least squares weights found in computing categorical regression.

Several studies have yielded varying results. Virtually without exception, increases in parallel forms reliability and internal consistency reliability have been reported when item-option weighted tests were compared to number-right scored or corrected-for-guessing scored tests.

The results concerning validity were not as consistent. Many studies reported losses in validity when item-option weights were applied to tests which were usually scored number-right or with the correction-for-guessing. However, other investigators reported no changes in validity due to the application of item-option weights.

This study investigated the effects of Guttman-type item-option weights on the reliability and validity of the Air Force Officer Qualifying Test (AFOQT-68) as used for the selection of pilots.

II. METHOD

The purposes of the analyses conducted in this study were: (a) to establish separate weights for the test item options of the AFOQT-68 Pilot Composite; (b) once established, to apply these weights to test data on an independent sample to ascertain the reliability of the composite scored with response option weights; (c) to estimate reliability (or stability) of the response weights by comparing weight sets established on two separate samples; and (d) to compare the validity of a conventionally derived Pilot Composite score with that of an item option weighted Pilot Composite score for prediction of graduation versus Flying Deficiency Elimination from undergraduate pilot training (UPT).

Between 1969 and 1972, 3,400 men were entered into UPT based on scores on the AFOQT-68's Pilot Composite. Two samples of 1,000 cases each were randomly selected from this pool. One of these was designated as a development sample and the other as a cross-application sample. The members of the cross-application sample were then reduced to only those who passed UPT and those who were eliminated for flying deficiency (N = 823). Traditionally, the Pilot Composite had been designed to predict Flying Deficiency Elimination rather than elimination due to academic or anxiety problems.

Form 68 of the AFOQT has 13 subtests. A complete description of the test and its development may be found in Miller (1968). The Pilot Composite is made up of seven subtests which may be classified as cognitive measures, perceptual measures, and a biographical inventory. Table 1 shows the subtest names, the number of items in each, and those subtests which comprise the Pilot Composite.

Several of the subtests are normally administered under speeded conditions with stringent time limits. It was necessary to reduce these speeded tests to only those items which at least 90 percent of the examinees reached. The technique used to develop the item-option weights required this change in test length because items which are infrequently answered would have spurious weights. No changes were made to the non-speeded subtests.

For all subjects, traditional correction-for-guessing scores were computed by the formula:

$$C = R - W/(K - 1)$$

where W is the number of items answered wrong and K is the number of item options. This is usually referred to as "formula scoring."

Table 1. **Name and Number of Items in the AFOQT-68 Subtests and Pilot Composite**

Subtest	Number of Items	Pilot Composite
Quantitative Aptitude	60	
Verbal Aptitude	60	
Officer Biographical Inventory	90[a]	
Scale Reading	26[a]	
Aerial Landmarks	34[a]	
General Science	24	
Mechanical Information	24	X
Mechanical Principles	24	X
Pilot Biographical Inventory	41[a]	X
Aviation Information	24	X
Visualization of Maneuvers	14[a]	X
Instrument Comprehension	14[a]	X
Stick and Rudder Orientation	48[b]	X

[a]Number of items reduced to eliminate speeded effect or multiple keyed responses for the purposes of this study.

[b]Original 24 double-response items broken into 48 single response items: 24 for stick and 24 for rudder for purposes of this study.

Weights for every item option were established on the development sample against an internal criterion. The weight for item 1, first option, was found by first identifying those subjects who selected that option. Their mean formula score on the *subtest*, with the item being considered deleted, was calculated and converted to a Z score. This procedure was carried out for every item option of every item. "Omit" was considered an item-option in all the subtests, and a weight was determined for this item-option for every item. Previously weighting was done across factorially complex scales or heterogeneous tests. In the present study, weights were developed using each subtest as a criterion. Both scores were transformed to standard scores for comparative purposes.

Statistical Approach

The validity was estimated by correlating the Pilot Composite scored by each method with the criterion variable. McNemar (1955) suggested a t-ratio for testing the difference between dependent correlations. It is

$$t = \frac{(r_{c1} - r_{c2})\sqrt{(N-3)(1+r_{12})}}{\sqrt{2(1 - r_{c1}^2 - r_{c2}^2 - r_{12}^2 + 2r_{c1}r_{c2}r_{12})}}$$

here 1 indicates the cross-validated item-option weighted scores, 2 indicates the formula scoring method, and C indicates the criterion. It is evaluated with N – 3 degrees of freedom. The difference between the validity of the corrected-for-guessing scored Pilot Composite (2) and that of the item-option weighted scored Pilot Composite (1) was tested by using this t-ratio.

The reliability for each of the subtests was estimated by calculating coefficient alpha (Cronbach, 1951). Wherry and Gaylord (1943) estimates of reliability were computed for the Pilot Composite.

III. RESULTS

Validity of the Pilot Composite

Table 2 shows the means and standard deviations of the AFOQT-68 subtests and Pilot Composite scored by the correction-for-guessing method and the means and standard deviations of the criterion

Table 2. **AFOQT-68 Subtest C Score Means and Standard Deviations (Development Sample N = 1,000 and Cross-Application Sample N = 823)**

Subtest	Sample: Development Mean	Development σ	Cross-Application Mean	Cross-Application σ
Quantitative Aptitude	.0309	1.0083	−.0060	.9975
Verbal Aptitude	−.0173	.9935	−.0214	.9959
Officer Biographical Inventory[a]	−.0248	.9900	−.0263	.9760
Scale Reading	.0389	.9929	.0144	.9959
Aerial Landmarks	−.0003	1.0071	−.0160	.9977
General Science	.0378	.9748	.0064	.9818
Mechanical Information	.0238	.9935	−.0061	.9980
Mechanical Principles	.0274	.9885	.0105	.9816
Pilot Biographical Inventory[a]	.0253	.9917	.0575	.9564
Aviation Information	.0121	.9990	−.0033	.9981
Visualization of Maneuvers	−.0071	.9642	−.0120	.9810
Instrument Comprehension	.0141	.9520	.0027	.9699
Stick and Rudder Orientation	.0263	.9951	.0243	.9975
Pilot Composite	.1274	4.0123	.0736	3.8212
Criterion Score	.8007	.3995	.7925	.4055

[a]Corrected-for-guessing scores were not calculated for biographical inventories, number-right scores have been substituted.

variable. The means and standard deviations of the AFOQT-68 subtests and the criterion are the same for both the development sample and the cross-application sample. Table 3 shows the means and standard deviations of the AFOQT-68 subtests and Pilot Composite scored by the application of item-option weights. The means and standard deviations of the item-option weighted subtests are the same for both groups.

Table 4 shows the validities of the Pilot Composite for predicting success in pilot training. A t-ratio was formed to test the difference between the validity of the formula score Pilot Composite scores and the validity of the cross-validated item-option scores. It was 2.9436 and was found to be significant at $P < .005$ ($Df = 820$).

Reliability

Pilot Composite and Subtests Reliability Table 5 presents the reliability for the subtests and Pilot Composite of the AFOQT-68 for the formula scores and the item-option weighted scores.

In all cases, the reliability of the item-option weighted scores is higher than the reliability of the formula scores.

Item-Option Weight Reliability The reliability of the item-option weights was estimated by a procedure suggested by Davis and Fifer (1959). An estimate of the stability of the weights was obtained by intercorrelating the weights found on the development sample with a set of weights developed on the cross-application sample. The reliability was .6028 ($F = 1563.5400$, $P < .005$).

IV. DISCUSSION

The results of this study cover three basic areas of concern about item-option weighted scores: validity, reliability, and item-option weight reliability.

Contrary to the results of some past research, this study has shown that the use of item-option weights does not necessarily lead to lowered test validity. Comparisons of unweighted score and weighted

Table 3. **AFOQT-68 Subtest Weighted Score Means and Standard Deviations (Development Sample N = 1,000 and Cross-Application Sample N = 823)**

Subtest	Sample: Development Mean	Development σ	Cross-Application Mean	Cross-Application σ
Quantitative Aptitude	.0333	1.0046	.0009	1.0018
Verbal Aptitude	−.0161	.9920	−.0188	.9961
Officer Biographical Inventory	.0257	.9911	−.0120	.9607
Scale Reading	.0309	.9927	.0169	.9956
Aerial Landmarks	−.0032	1.0064	−.0164	.9982
General Science	.0437	.9804	.0222	.9725
Mechanical Information	.0297	1.0033	.0003	1.0080
Mechanical Principles	.0328	.9943	.0160	.9884
Pilot Biographical Inventory	.0217	.9991	.0630	.9713
Aviation Information	.0157	1.0046	.0141	1.0020
Visualization of Maneuvers	.0082	.9579	.0047	.9950
Instrument Comprehension	.0113	.9617	.0003	.9691
Stick and Rudder Orientation	.0295	.9935	.0273	.9929
Pilot Comprehension	.1490	4.1079	.1257	3.8976

Table 4. **Validity of AFOQT-68 Subtests and Pilot Composite for Predicting Success in Undergraduate Pilot Training in the Cross-Application Sample (N = 823)**

Subtest	Validity: Formula Scores	Item-Option Weighted Scores
Quantitative Aptitude	.0249	.0451
Verbal Aptitude	−.0944	−.0959
Officer Biographical Inventory	−.0251[a]	.0236
Scale Reading	.0994	.1137
Aerial Landmarks	.0442	.0402
General Science	.0593	.0649
Mechanical Information	.0027	.0221
Mechanical Principles	.0221	.0280
Pilot Biographical Inventory	.1223[a]	.1383
Aviation Information	.0570	.0569
Visualization of Maneuvers	.0613	.0901
Instrument Comprehension	.1368	.1457
Stick and Rudder Orientation	.1123	.1608
Pilot Composite	.1317	.1621

[a]C scores were not computed for biographical inventories. Number-right scores have been substituted.

Table 5. **The Reliability of the Formula and Cross-Validated Item-Option Weighted Scores for the Subtests and Pilot Composite of the AFOQT-68 for the Cross-Application Sample**

Subtest	Reliability	
	Formula Scores	Item-Option Weighted Scores
Quantitative Aptitude	.8165	.8363
Verbal Aptitude	.7830	.8462
Officer Biographical Inventory	.6568[a]	.9109
Scale Reading	.5504	.6663
Aerial Landmarks	.7899	.8041
General Science	.7534	.7815
Mechanical Information	.6727	.7117
Mechanical Principles	.6636	.6834
Pilot Biographical Inventory	.6796[a]	.8430
Aviation Information	.6964	.7576
Visualization of Maneuvers	.5236	.6721
Instrument Comprehension	.6133	.7050
Stick and Rudder Orientation	.8096	.8458
Pilot Composite[b]	.8543	.8839

[a]C scores were not computed for biographical inventories. Number-right scores have been substituted.

[b]Coefficient alpha was used to estimate reliability for the subtests. Reliability for the Pilot Composite was estimated by the Wherry and Gaylord method.

score validities in Table 4 indicate that, with few exceptions, validity can be maintained or increased by the use of item-option weights.

The exact cause for the maintenance or increases in validity is not known. It may be that weights developed per subtest do not adversely effect the factorial heterogeneity of the Pilot Composite. Guttman's original method was specifically designed for use with univocal tests. Many of the subtests are very complex as evidenced by their low internal consistency (see Table 5). Visualization of maneuvers is an example of this heterogeneity with an alpha coefficient of .5236 unweighted. Item-option weighting only increased this to .6721.

Internal consistency is a function of item intercorrelations (Cronbach, 1951). Highly consistent tests have high item-intercorrelations. However, it is clear from the Wilks (1938) proof that the effects of weights applied to a large set of positively intercorrelated variables are inconsequential. Applying this model to item-option categories would lead to the conclusion that tests with lower internal consistency should benefit from item-option weighting more than tests with high internal consistency.

The reliability of subtests or composites as measured by calculating coefficient alpha or by the Wherry and Gaylord method increased as expected. The two biographical inventories showed the largest reliability gains while quantitative aptitude and verbal aptitude tests, similar to those used in several other studies, showed very minor gains in reliability. The Pilot Biographical Inventory also showed a substantial gain in validity. It may be that the use of item-option weights with biographical inventories leads to higher reliability and higher validity. Torgerson (1958) classified Guttman-type item-option weights as a scaling technique. Item-option weighting of biographical inventories, even those with a priori weights, may be closer to the use originally intended for the Guttman technique.

The reliability of the item-option weights was moderate, as expected. However, the keyed item options showed a reliability of .891.

In this study, there was little omitting of items. In some cases, as few as 6 in 1,000 subjects omitted an item. This led to a large standard error of the mean and, thus, to unreliable weights for the "omit" item option. The keyed answers were generally selected by the largest group of examinees which reduced the standard error of the mean and produced highly reliable item-option weights.

Although the reliability of the overall set of item-option weights was only moderate, the most frequently selected item options had the most reliable item-option weights.

V. CONCLUSIONS

The use of Guttman-type item-option weights was effective in increasing both subtest and Pilot Composite reliability, as well as Pilot Composite validity for predicting UPT success versus elimination for reason of flying deficiency. The set of item-option weights was found to be moderately reliable. Item-option weights for the keyed item options were found to be highly reliable.

In order to make use of the item-option weighting procedure, machine scoring of tests at recruiting stations would be mandatory. The establishment of item-option weights on appropriate samples would also have to become a regular part of the test development and revision process.

REFERENCES

Cronbach, L.J. Coefficient alpha and the internal structure of tests. *Psychometrika,* 1951, **16**, 297–334.

Davis, F.B., & Fifer, G. The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. *Educational and Psychological Measurement*, 1959, **19**, 159–170.

Guttman, L. The quantification of a class of attributes: A theory and method of scale construction. In M. P. Horst et al. (Eds.), *The prediction of personal adjustment*. New York: Social Science Research Council, 1941.

McNemar, Q. *Psychological statistics* (2nd ed.). New York: John Wiley and Sons, 1955, p. 124.

Miller, R. *Development of officer selection and classification tests–1968.* AFHRL-TR-68-104, AD-679 989. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, July 1968.

Torgerson, W.A. *Theory and methods of scaling.* New York: John Wiley and Sons, Inc., 1958, 338–345.

Wherry, R., & Gaylord, R. The concept of test and item reliability in relation to factor pattern. *Psychometrika,* 1943, **8**, 247–264.

Wilks, S.S. Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika,* 1938, **3**, 23–40.

☆U.S. GOVERNMENT PRINTING OFFICE: 1977-771-057/26